doi:10.7515/JEE201405006

长江水系水质综合评价、预测的投影寻踪建模与实证研究

楼文高[1,2]，熊　聘[2]
（1. 上海商学院，上海 200235；2. 上海理工大学 光电信息与计算机工程学院，上海 200093）

摘　要：根据我国地表水环境质量标准和长江水系 2006—2010 年 103 个国控断面 8 个水质监测指标的数据，建立了长江水系水质综合评价的投影寻踪分类（PPC）模型。研究表明：在国控断面上，长江水系整体水质较好，Ⅰ～Ⅲ类水质占 92.82%，其中Ⅰ类和Ⅱ类水质分别占 62.52% 和 20.58%；上游水质好于下游水质；在化工工业较发达地区，Hg、石油类、挥发酚和氨氮的污染较严重，出现了劣Ⅴ类水质；与 2009 年相比，2010 年水质出现了恶化。投影窗宽半径 R 值对 PPC 建模结果有显著影响，R 取较小值方案（R=0.1S_z）和较大值方案（$r_{max} \leqslant R \leqslant 2p$）都是不合理的，取中间适度值方案（$r_{max}/5 \leqslant R \leqslant r_{max}/3$）才是合理和正确的。

关键词：长江水系；水质综合评价；投影寻踪建模；投影窗宽半径 R 值
中图分类号：X824；TP391　　**文献标志码**：A　　**文章编号**：1674-9901(2014)05-0344-09

Water quality comprehensive evaluation and prediction of the Yangtze River applying projection pursuit clustering technique and its positive analysis

LOU Wen-gao[1,2], XIONG Pin[2]
(1. Shanghai Business School, Shanghai 200235, China; 2. School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: According to the environmental quality standards for surface water and the measured data of eight indexes describing surface-water quality of the 103 government-control sections of the Yangtze River during the years from 2006 to 2010, the water quality evaluation model applying projection pursuit clustering (PPC) technique was established in this paper. The cases study results show that the water quality on the control sections is quite good, the type Ⅰ～Ⅲ taking up 92.82%, of which type Ⅰ and type Ⅱ taking up 62.52% and 20.58%, respectively. The water quality in upper region is better than that in the downstream. The water pollution caused by Hg, petroleum, volatile phenol and NH_3-N is serious in the chemical-industry developed areas, and the water quality is type Ⅴ. Comparison with that in year 2009, the water pollution in year 2010 is more serious. The cutoff radius R value (CRRV) obviously determines the results of PPC model. The theoretical analyses and positive researches show that the strategy of taking the smaller CRRV (R=0.1S_z) or the larger CRRV ($r_{max} \leqslant R \leqslant 2p$) is absolutely unreasonable and unfeasible, and the strategy of taking the moderate-suitable CRRV ($r_{max}/5 \leqslant R \leqslant r_{max}/3$), however, is reasonable and effective.

Key words: the Yangtze River; water quality comprehensive evaluation; projection pursuit clustering (PPC) modelling; cutoff radius R value

收稿日期：2014-07-13
基金项目：上海高校知识服务平台“上海商贸服务业知识服务中心”建设项目；上海市重点学科“商务经济学”项目
通讯作者：熊　聘，E-mail: pinxiongcn@foxmail.com

长江水系为我国提供了近 40% 的水能资源，在国民经济发展中占有举足轻重的战略地位。然而，由于环境的严重污染导致长江水系水质污染也日益严重。为了对长江水系水资源进行充分利用、管理和正确规划提供科学的决策依据，必须对水系水质进行正确评价与预测。

水质综合评价方法有指数评价法、集对分析法、物元分析法、模糊数学法、灰色评价法、Logistic 模型、主成分分析（PCA）、支持向量机（SVM）、人工神经网络（ANN）模型及其与 ANN 的组合评价等很多种方法（李惠明和尚广萍，1991；李祚泳等，2001；金菊良等，2003；楼文高和王延政，2003；陈润羊等，2008；刘小楠和崔巍，2009），这些方法各具特色与缺陷，如前五种方法必须首先由专家法和信息熵等主客观方法确定权重，结果的合理性随权重的合理性而改变，属于半定量研究方法；PCA 方法必须满足大样本条件（MacCallum et al，1999），而且，样本不同，其评价结果也随之改变；ANN 建模过程十分复杂，必须判定多种现象和确定多个合理参数，只有在遵循基本原则的前提下才可能取得可靠的结果（楼文高和王延政，2003；楼文高和乔龙，2011；StatSoft Inc，2011）；SVM 模型须人为判定两个参数的合理值，等等。另一方面，Friedman and Tukey（1974）提出的可用于高维、非线性、非正态分布数据建模的投影寻踪分类（PPC）技术，自张欣莉等（2000）将其应用于水质评价以来，获得了广泛的应用（王顺久等，2002；封志明等，2005；付强和赵小勇，2006；楼文高和乔龙，2013）。但是，关于如何确定合理的窗宽半径 R 值以及判定最优化算法是否求得了真正的全局最优解，应用中也出现了不少错误和荒谬的结果。为此，考虑到数据的公开性和可获得性，本文根据《中国环境年鉴》中长江水系 103 个国控监测断面 2006—2010 年的主要水质指标监测数据，选取地表水环境质量标准（GB3832-2002）中 DO、COD_{Mn}、Hg 和 NH_3-N 等 8 种主要污染物指标，采用 PPC 建模技术，综合评价研究长江水系所有国控断面的水质情况。研究结果表明：长江水系水质整体上以Ⅰ类（占 62.52%）、Ⅱ类（占 20.58%）和Ⅲ类（占 9.71%）水质为主，但个别断面也出现了劣Ⅴ类水质的情况。并且研究了不同 R 值对建模结果的影响以及出现的明显错误，进而提出了选取 R 值的合理范围。

1　水质综合评价 PPC 建模原理简介

1.1　长江水系水质监测指标及其评价标准

根据《中国环境年鉴》（2007—2011 年）列出的长江水系国控断面 9 个主要监测指标 2006—2010 年的年均值，因为Ⅰ～Ⅴ级水质的 pH 值均为 6~9，故我们选取 DO、COD_{Mn} 和 NH_3-N 等 8 个主要监测指标来综合评价长江水系的水质。我国地表水环境质量标准（GB3832-2002）如表 1 所示，同时，根据五级水质的指标值及其变化规律或者趋势以及长江水系水质指标实测值的范围确定了各指标的最大值和最小值（也列于表 1 中）。

表 1　地表水环境质量标准（GB3838-2002）
Table 1　Environmental quality standards for surface water

评价指标		Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ	最大值	最小值
DO	≥	7.5	6	5	3	2	10.6	0.5
COD_{Mn}	≤	2	4	6	10	15	20	0.5
BOD_5	≤	3	3	4	6	10	15	0.5
NH_3-N	≤	0.15	0.5	1	1.5	2	5	0.01
Hg	≤	0.00005	0.00005	0.0001	0.001	0.001	0.35	0.00001
Pb	≤	0.01	0.01	0.05	0.05	0.1	0.15	0.0003
挥发酚	≤	0.002	0.002	0.005	0.01	0.1	1	0.001
石油类	≤	0.05	0.05	0.05	0.5	1	1.5	0.001

1.2　水质综合评价 PPC 建模原理简介

根据表 1 的水质评价标准，可用由 Friedman and Tukey（1974）提出并被张欣莉等（2000）首先采用、后被国内其他学者广泛采用的一维 PPC

模型（王顺久等，2002；封志明等，2005；付强和赵小勇，2006；楼文高和乔龙，2013）建立长江水系水质综合评价模型。Friedman 等提出“应使样本投影点整体上尽可能分散，局部上尽可能密集”（Friedman and Tukey，1974；张欣莉等，2000），其目标函数为使样本投影值标准差 S_z 与局部密度值 D_z 乘积的最大化，即：

$$Q(a)=\max(S_z\times D_z)$$
$$s.t.\sum_{j=1}^{p}a_j^2=1,\ 1\geqslant a_j\geqslant -1 \quad (1)$$

其中表示投影点整体上离散程度的样本投影值标准差 $S_z=\sqrt{\{\sum_{i=1}^{n}[z(i)-E(z)]^2\}/(n-1)}$（即 S_z 越大表示投影点整体上越分散），投影点局部密集程度的局部密度值 $D_z=\sum_{i=1}^{n}\sum_{k=1}^{n}[R-r_{i,k}]\cdot u[R-r_{i,k}]$（$D_z$ 越大表示投影点局部越密集），其他符号和公式的意义详见文献 Friedman and Tukey（1974）、张欣莉等（2000）、楼文高和乔龙（2013）。

由上式可知，投影密度窗口半径 R 值显著影响其最优化结果 —— 最佳投影向量以及样本投影值，楼文高和乔龙（2013）深入研究了 R 值的本质及其对建模结果的影响，进而提出了选取 R 值的合理范围应该为 $r_{max}/5 \leqslant R \leqslant r_{max}/3$，即平均有 1/5~1/3 的样本点在投影窗宽内。由于（1）式为高维、复杂的非线性不等式约束最优化问题，很难求得真正的全局最优解。笔者采用群搜索算法（GSO）对（1）式进行最优化求解。

2 建立水质综合评价 PPC 模型及其长江水系的水质与变化趋势

为了消除水质指标不同量纲对建模结果的影响，本文对表 1 的各指标值进行开放性更好的（即实际样本的最大值和最小值可以大于或者小于表 1 所示的最大值和最小值）零均值归一化（即均值为 0，方差为 1）处理，而对 DO 先进行正向化处理，即所有指标的归一化值越大，相应的投影值也越大，其水质就越差。把各样本归一化数据导入笔者研发的基于 GSO 算法的 PPC 程序，得到了 $R=r_{max}/5$ 时的 PPC 建模结果，即最佳投影向量 $a_{1\sim8}$=（0.143，0.252，0.336，0.367，0.488，0.287，0.484，0.338），各样本的投影值 $z(1)\sim z(7)$=（−2.102，−1.643，−1.436，−0.938，−0.205，1.052，5.272），样本投影值标准差 S_z=2.550，局部密度值 D_z=23.102，目标函数值 $Q(a)$=58.899，投影窗宽半径 R=1.475，样本之间的最大距离 r_{max}=7.374。由五个标准样本的意义可知，Ⅰ～劣Ⅴ类水质的投影值范围分别为小于等于 −1.643、（−1.643，−1.436]、（−1.436，−0.938]、（−0.938，−0.205]、（−0.205，1.052] 和大于 1.052。

对《中国环境年鉴》给出的长江水系 103 个国控断面的 8 个主要监测指标值，也进行相同的零均值归一化处理，再代入上述 PPC 模型，就得到了各个国控断面 2006—2010 年的水质投影值 $z(i)$，根据上述各级不同水质的投影值范围，就可以很便捷地判定各断面的综合水质类别。为节省篇幅，表 2 仅列出各省市交界的典型国控断面和苏州、上海等合计 22 个断面的 PPC 模型投影值及其综合评价的水质类别。

由研究结果和表 2 可知，长江水系国控断面的水质呈以下特征：（1）综合评价水质以Ⅰ类和Ⅱ类为主，在 515 个样本（103 个断面、五年）中，Ⅰ类水质 322 个，占 62.52%，Ⅱ类水质 106 个，占 20.58%；（2）2006—2010 年分别有 37 个、24 个、17 个、10 个和 15 个断面的综合水质最差，说明整体上长江水系水质有所改善，沿长江流域各省市或者城市的水污染治理工程发挥了重要作用，基本遏止了水质恶化的趋势，如沿河断面、鲢鱼溪断面的水质从Ⅱ类变为Ⅰ类，新甸铺断面水质从Ⅳ类变为Ⅲ类，吴淞口水质从Ⅲ类变成Ⅱ类；但与 2009 年相比，2010 年水质整体上有所恶化；（3）上游水质要好于下游水质，流经大城市的水质相对较差（如苏州、上海等）；（4）流经安徽、四川、江苏和湖北等省市后的综合水质明显变差，这些省市必须采取切实有效措施，进一步加大水污染治理力度，彻底扭转这种状况。

3 结果与讨论

3.1 各水质评价指标重要性分析

根据 PPC 建模的最优化结果，8 个评价指标都是重要的，但从重要性来看，Hg 最重要，其次是挥发酚，然后是 NH_3-N、石油类、BOD_5、Pb、COD_{Mn} 等，DO 最不重要，权重最大值与最小值之比达到了 3.4 倍，说明它们的差异显著。因此，从整体上有效改善长江水系综合水质的角度来看，首先应显著减少 Hg 和挥发酚的排放，其次是明显降低石油类、氨氮、BOD_5 和 Pb 等的排放，这样才能起到事半功倍的效果，否则，往往事倍功半，起不到明显的效果。

表 2　长江水系典型国控断面水质投影值、综合评价结果及其类别

Table 2　The projected values and the type of the water quality in the typical state-controlled sections of Yangtze River

断面代码	地区名称	断面名称	三峡库区	省界断面	2006		2007		2008		2009		2010	
					PPC 值	类别	PPC 值	类别	PPC 值	类别	PPC 值	类别	PPC 值	类别
1	攀枝花	龙洞	上游区	滇 — 川	−1.875	Ⅰ	−1.833	Ⅰ	−1.764	Ⅰ	−1.774	Ⅰ	−1.798	Ⅰ
2	水富	铁路桥	上游区	滇 — 川	−1.656	Ⅰ	−1.856	Ⅰ	−1.833	Ⅰ	−1.870	Ⅰ	−1.872	Ⅰ
3	重庆	朱沱	影响区	川 — 渝	−1.486	Ⅱ	−1.525	Ⅱ	−1.494	Ⅱ	−1.604	Ⅱ	−1.704	Ⅰ
4	重庆	培石	库区	渝 — 鄂	−1.719	Ⅰ	−1.739	Ⅰ	−1.688	Ⅰ	−1.674	Ⅰ	−1.756	Ⅰ
5	岳阳	城陵矶		湘 — 鄂	−1.675	Ⅰ	−1.703	Ⅰ	−1.735	Ⅰ	−1.685	Ⅰ	−1.676	Ⅰ
6	九江	姚港		赣 — 鄂	−1.784	Ⅰ	−1.729	Ⅰ	−1.747	Ⅰ	−1.770	Ⅰ	−1.758	Ⅰ
7	安庆	皖河口		赣 — 皖	−1.760	Ⅰ	−1.766	Ⅰ	−1.796	Ⅰ	−1.812	Ⅰ	−1.783	Ⅰ
8	南京	江宁河口		皖 — 苏	−1.785	Ⅰ	−1.708	Ⅰ	−1.768	Ⅰ	−1.781	Ⅰ	−1.767	Ⅰ
9	南通	姚港		苏 — 沪	−1.703	Ⅰ	−1.709	Ⅰ	−1.743	Ⅰ	−1.769	Ⅰ	−1.787	Ⅰ
10	遂宁	老池	上游区	川 — 渝	−1.687	Ⅰ	−1.781	Ⅰ	−1.696	Ⅰ	−1.631	Ⅱ	−1.727	Ⅰ
11	广元	八庙沟	上游区	甘 — 川	−1.702	Ⅰ	−1.888	Ⅰ	−1.870	Ⅰ	−1.873	Ⅰ	−1.828	Ⅰ
12	岳池	赛龙乡	上游区	川 — 渝	−1.546	Ⅱ	−1.664	Ⅰ	−1.611	Ⅱ	−1.670	Ⅰ	−1.684	Ⅰ
13	铜仁	沿河	上游区	黔 — 渝	−1.628	Ⅱ	−1.765	Ⅰ	−1.745	Ⅰ	−1.726	Ⅰ	−1.821	Ⅰ
14	重庆	利泽	影响区	川 — 渝	−1.705	Ⅰ	−1.703	Ⅰ	−1.690	Ⅰ	−1.745	Ⅰ	−1.743	Ⅰ
15	赤水	鲢鱼溪	影响区	黔 — 川	−1.640	Ⅱ	−1.550	Ⅱ	−1.765	Ⅰ	−1.769	Ⅰ	−1.831	Ⅰ
16	武都	绸子坝		甘 — 川	−1.847	Ⅰ	−1.844	Ⅰ	−1.793	Ⅰ	−1.825	Ⅰ	−1.841	Ⅰ
17	十堰	羊尾		陕 — 鄂	−1.807	Ⅰ	−1.855	Ⅰ	−1.792	Ⅰ	−1.795	Ⅰ	−1.681	Ⅰ
18	滁州	汊河		皖 — 苏	0.487	Ⅴ	0.033	Ⅴ	−0.234	Ⅳ	0.644	Ⅴ	−1.003	Ⅲ
19	南阳	梅湾		豫 — 鄂	−1.450	Ⅱ	−1.459	Ⅱ	−1.574	Ⅱ	−1.421	Ⅲ	−1.468	Ⅱ
20	南阳	新甸铺		豫 — 鄂	−0.723	Ⅳ	−0.865	Ⅳ	−0.951	Ⅲ	−1.158	Ⅲ	−1.343	Ⅲ
21	苏州	轻化仓库			0.484	Ⅴ	−0.220	Ⅳ	−0.696	Ⅳ	−0.671	Ⅳ	−0.892	Ⅳ
22	上海	吴淞口			−1.109	Ⅲ	−1.227	Ⅲ	−1.143	Ⅲ	−0.976	Ⅲ	−1.641	Ⅱ

3.2　长江水系整体水质情况及其变化趋势分析

根据 103 个国控断面 2006—2010 年的水质综合评价结果可知，长江水系整体水质呈如下特征：（1）长江水系综合水质处于良好状态，以Ⅰ类和Ⅱ类水质为主，其中Ⅰ类、Ⅱ类和Ⅲ类分别占 62.52%、20.58% 和 9.71%，而Ⅳ和Ⅴ类分别仅占 4.08% 和 1.75%，劣Ⅴ类占 1.36%。（2）介于省市分界的国控断面综合水质有所好转，说明各省市都十分重视水污染治理工程，并取得了显著效果；与此同时，在 103 个国控断面上，虽然 2007—2009 年期间水质有所好转，但 2010 年的水质差于 2009 年（如：衡阳从Ⅱ类水质变成Ⅲ类，六盘山从Ⅱ类变成Ⅴ类）。这是否意味着各省市放松了对水环境污染的治理？值得我们反思和警惕。（3）上游地区（断面）水质好于下游地区，大城市（断面）水质比其他区域水质差。（4）长江水系在贵州省、四川省、江西省、安徽省、湖北省和江苏省的水质较差，应该是长江水系水环境污染治理的重点。由于这些区域酿酒业、造纸业、皮革加工业等化工工业发展迅速，工业污水虽经处理，但水体中 Hg、石油类、挥发酚和氨氮等的含量仍然较高，导致水质大多低于Ⅲ类，有些断面甚至出现了Ⅴ类和劣Ⅴ类水质，水污染情况十分严重。

3.3　窗宽半径 *R* 值对 PPC 建模结果和长江水系水质评价结果的影响

R 值是决定 PPC 模型最优化结果的唯一参数，但不同的学者提出了多种不同的选取 *R* 值的方案：（1）Friedman and Tukey（1974）提出较小值方案，即 $R=0.1S_z$ 或者更小的 $R=(0.01\sim0.001)S_z$（张欣莉等，2000；封志明等，2005）；（2）王顺久等（2002）提出取较大值方案，取 $r_{max} \leqslant R \leqslant 2p$，通常取 $R=p$；（3）楼文高和乔龙（2013）提出取中间适度值方案，通常取 $r_{max}/5 \leqslant R \leqslant r_{max}/3$。

取较小值方案时，窗宽半径内只包含很少的样本点，取较大值方案时，窗宽半径内包含了所

有的样本点，取中间适度值方案，窗宽半径内样本点既不太多，也不太少。那么，三种选取 R 值的不同方案将对 PPC 建模结果产生怎样的影响呢？到底哪个方案更合理？建模结果更可靠？

为了分析 R 值对 PPC 建模结果的影响规律，笔者分别取 $R \leqslant 0.001S_z$、$0.01S_z$、$0.05S_z$、$0.1S_z$、$0.25S_z$、$0.5S_z$、$r_{max}/5$、$r_{max}/4$、$r_{max}/3$、S_z、$r_{max}/2$、r_{max} 和 $R=p=8$ 建模，表 3 所示为取不同 R 值时的 PPC 建模结果。分析上述建模结果，可以得到如下规律：（1）$R \in [0.001S_z, 0.25S_z]$ 时，最佳投影向量发生了很明显的改变，且没有规律性，而且部分指标的性质是错误的（即表 3 中权重小于 0 的情况，下同），如 a_1 从 0.328 变为 −0.328、0.057、−0.013、−0.024，而且最优化过程往往无法求得真正的全局最优解，针对每个 R 值，笔者通过调整不同的参数组合，应用群搜索算法或者混沌 ABC 算法计算 30 余次，表 3 所示是计算所得的最好结果，但还不是真正的全局最优解；（2）$R \leqslant 0.001S_z$ 或 $R \geqslant r_{max}$ 时，最优化结果很稳定，而且基本相同，在此范围内，R 值越小或者越大，各指标权重变化越小；（3）$R \in [0.5S_z, r_{max}/2]$ 时，最优化过程很容易求得真正的全局最优解，所有指标的性质都是正确的，最佳投影向量和各样本投影值随 R 值变化而改变，但相差不是很大，体现了 PPC 建模可以揭示出高维数据不同结构特点的精髓。

此外，还有不少文献取 R 为常数，为此笔者取 $R=10^{-5}$、10^{-4}、10^{-3} 以及与表 3 基本对应的 R 值（如 0.0027、0.068、0.138、0.360、1.260、1.470、1.900、2.500、3.800、7.800）、10 和 100 分别建模，表 4 所示是其 PPC 建模结果。由表 4 可知，建模结果的规律与 R 取非常数时基本一致，但也表现出不同的特点：（1）$R \in [10^{-5}, 1.900]$（1.900 相当于 $r_{max}/2.63$ 或者 $1.04S_z$）时，部分指标的权重出现了小于 0 的情况，如 $a_1 \leqslant 0$ 的情况（用波浪线表示小于 0，下划线表示几乎为 0，下同），这些指标的性质肯定是错误的。此外，出现了多个指标的权重同时为 0 的情况（如 $R=0.0027$ 时），这种情况肯定也是错误的；（2）$R \in [10^{-4}, 0.068]$ 时，出现了多个样本值相等的情况，事实上，这些样本值是区分不同类别水质的分界值，如果都相等，无法区分不同水质等级，所以也肯定是不合理的；（3）虽然 R 大于 3.800（相当于 $r_{max}/2$）时，最优化过程能求得真正的全局最优解，但 $R=3.800$ 时 a_1 的权重仅为 0.036，明显小于其实际值，也是不合理的。$R=7.800$（此时 R 已大于 r_{max}）的权重与非常数时 $R=r_{max}/3$ 的权重基本相当；（4）当 $R \leqslant 10^{-5}$ 或 $R \geqslant 50$ 时，各个指标的权重才基本稳定不变。

因此，R 取常数时第（2）~（4）个规律与取非常数时的规律不同。而且，即使 R 值基本相同，取常数与非常数的建模结果（权重等）却相差很大。为此，笔者认为 R 为常数也是不合理的。

3.4 不同群智能最优化算法的性能比较

笔者用多种改进 PSO（IPSO）、多种遗传算法（GA）、基于实数编码的加速遗传算法（RAGA）（该最优化算法已被很多学者用于 PPC 建模）、多智能体遗传算法（MGA）、人群搜索算法（SOA）、遗传 -PSO 组合算法（GA-PSO）、鸡群算法（CSO）、果蝇算法（FOA）、人工鱼群算法（AFSA）、萤火虫算法（FFA）、布谷鸟搜索算法（C-SO）、蚁群算法（ACO）、蚁狮算法（ALO）、差分算法（DA）、和声搜索算法（HS）、灰狼搜索算法（GWO）、资源竞争算法（COR）、蜘蛛算法（SSO）、花授粉算法（FPA）、细菌算法（BFO）、重力搜索算法（GSA）、群搜索算法（GSO）和人工蜂群算法（ABC）等对 PPC 模型进行最优化求解。针对每种算法，通过调整不同的参数组合，平均计算 30 次以上，研究发现：（1）在 R 取中间适度值和较大值方案时，绝大多数群智能算法都基本能求得真正的全局最优解，但收敛速度和收敛性能有较大差异，其中 ABC、GSO、GSA、GWO、ALO、SOA、IPSO 和 CSO 等的收敛速度较快，也基本每次都能求得真正的全局最优解，而其他方法则往往收敛较慢，也不能保证每次都求得真正的全局最优解；（2）R 取较小值方案时（如 $R \in [0.001S_z, 0.25S_z]$ 或者 $R \in [10^{-4}, 0.360]$），上述群智能最优化算法往往都不能求得真正的全局最优解，但是，在接近全局最优解邻域范围内，GSO、GSA、ABC 和 MGA 的收敛性能较好，收敛速度也较快，应优先推荐使用，而其他方法，无论是收敛性能还是收敛速度，都不太理想；（3）被很多学者用于 PPC 建模的 RAGA 和 ACO 等算法，其收敛性能和收敛速度都明显差于 ABC、IPSO、GSO、GSA 和 SOA 等算法。

为了从理论上进一步研究上述实证研究结果所表现出来的规律性，笔者分析讨论如下：

当 $i=k$ 时，（1）式的 $r_{i,k}=r_{k,i}=0$，则 $D_z=\sum_{i=1}^{n}\sum_{k=1}^{n}[R-r_{i,k}]\cdot u(R-r_{i,k})=nR+2\sum_{i=1}^{n}\sum_{k=i+1}^{n}(R-r_{i,k})\cdot u(R-r_{i,k})$。当 R 取较小值方案时，往往仅有少数样本点能满足 $R>r_{i,k}(i\neq k)$ 条件（假设 m 个）或者出现多个样本值相同的情况，

此时 $r_{i,k}=0(i\neq k)$，则 $D_z=(2m+n)R-2\sum_{l=1}^{m}r_l$，即目标函数 $Q(a)=\max[0.1\times(2m+n)\times S_z^2-2S_z\sum_{l=1}^{m}r_l]$，也就是说，目标函数值主要取决于 S_z 的最大化。而且，由于 $S_z\sum_{l=1}^{m}r_l$ 的值随各评价指标权重的改变而改变，目标函数存在很多的局部极小点，导致最优化过程往往收敛于局部极小点（如表 3 和表 4 所示），而无法求得真正的全局最优解。如果 R 值非常小，而且最优化算法设置的收敛精度如果不是很高（如为 10^{-4}，实践上收敛精度已经足够高，但目标函数值已很小），则绝大部分甚至所有样本点均满足 $R<r_{i,k}(i\neq k)$ 条件，则 $Q(a)=\max[0.1\times nS_z^2]$，此时，目标函数值 $Q(a)$ 值只与类间距离 S_z 有关，其最优化收敛过程也必定是稳定的，可以求得真正的全局最优解。

如果 R 采用较大值方案（$r_{\max}\leqslant R\leqslant 2p$），则对所有样本点都有 $R\geqslant r_{\max}\geqslant r_{ik}$，$D_z=n^2R-2\sum_{i=1}^{n}\sum_{k=i+1}^{n}r_{i,k}$，则目标函数 $Q(a)=\max[S_z\times(n^2R-2\sum_{i=1}^{n}\sum_{k=i+1}^{n}r_{i,k})]$。因此，只有同时使 S_z 最大化（即使所有样本点尽可能分散）和 $r_{i,k}$ 最小化（即使所有样本点尽可能密集），目标函数 $Q(a)$ 才可能取得最大值，而所有样本的 S_z 最大化和 $r_{i,k}$ 最小化显然是相互矛盾的。实践中，当 $R\geqslant 2r_{\max}$ 时就有 $n^2R>>2\sum_{i=1}^{n}\sum_{k=i+1}^{n}r_{i,k}$，即 $Q(a)=\max(n^2RS_z)$，也就是说，此时只要使 S_z 最大化，$Q(a)$ 就取得最大值。

从上述实证结果和理论分析可知，R 取很小的值和较大值方案时，PPC 建模结果必定是相同的，而且都只使 S_z 最大化，即仅仅达到了“使样本投影点整体上尽可能分散”的目标，而没有同时实现“使样本点局部尽可能密集”的目标，没有完全实现 Friedman and Tukey（1974）提出的 PPC 建模的目标，是不合理的。而且，R 取较小值方案时，R 值不同，PPC 建模结果也不同，有时会出现很大的变化，也没有规律性可循。另一方面，R 取中间适度值方案时，目标函数完整体现了 Friedman 等提出的 PPC 建模基本思想——“样本投影点整体上尽可能分散，局部尽可能密集”的要求，所以是合理和正确的。

表 3　窗宽半径 取不同值（三种方案）时的 PPC 建模结果对比

Table 3　Comparison of PPC model in different cutoff radius R (three methods)

最优化结果		$0.001S_z$	$0.01S_z$	$0.05S_z$	$0.1S_z$	$0.25S_z$	$0.5S_z$	$r_{\max}/5$	$r_{\max}/4$	$r_{\max}/3$	S_z	$r_{\max}/2$	$r_{\max}$	8
样本投影值	$z(1)$	−2.536	−1.274	−0.514	−0.518	−0.556	−2.000	−2.102	−2.262	−2.307	−2.271	−2.364	−2.521	−2.373
	$z(2)$	−1.854	−1.273	−0.513	−0.536	−0.563	−1.615	−1.643	−1.725	−1.745	−1.742	−1.767	−1.837	−1.787
	$z(3)$	−1.536	−1.273	−0.497	−0.518	−0.558	−1.446	−1.436	−1.479	−1.485	−1.497	−1.487	−1.516	−1.514
	$z(4)$	−0.888	−0.845	−0.512	−0.518	−0.556	−0.962	−0.938	−0.930	−0.915	−0.930	−0.904	−0.882	−0.919
	$z(5)$	0.045	−0.426	−0.52	−0.536	−0.556	−0.238	−0.205	−0.130	−0.109	−0.124	−0.077	0.026	−0.067
	$z(6)$	1.520	0.700	−0.513	−0.518	−0.470	1.054	1.052	1.202	1.230	1.238	1.269	1.457	1.333
	$z(7)$	5.249	4.392	3.069	3.144	3.258	5.207	5.272	5.324	5.33	5.324	5.329	5.274	5.327
最佳投影向量系数	a_1	0.328	−0.328	0.057	−0.013	−0.024	0.054	0.143	0.208	0.234	0.193	0.272	0.343	0.245
	a_2	0.363	0.211	−0.242	−0.162	−0.138	0.265	0.252	0.286	0.293	0.302	0.302	0.343	0.323
	a_3	0.371	0.23	−0.028	0.035	0.067	0.328	0.336	0.350	0.348	0.343	0.349	0.369	0.353
	a_4	0.371	0.409	0.286	0.337	0.269	0.356	0.367	0.372	0.379	0.378	0.383	0.374	0.379
	a_5	0.324	0.409	0.701	0.734	0.721	0.466	0.488	0.446	0.438	0.429	0.427	0.355	0.400
	a_6	0.366	0.407	−0.064	−0.084	−0.068	0.303	0.287	0.312	0.325	0.336	0.324	0.349	0.343
	a_7	0.338	0.409	0.601	0.556	0.615	0.478	0.484	0.446	0.437	0.434	0.423	0.360	0.405
	a_8	0.364	0.356	0.013	−0.059	−0.036	0.394	0.338	0.344	0.326	0.354	0.318	0.334	0.355
目标函数	S_z	2.673	2.063	1.353	1.3864	1.437	2.514	2.550	2.612	2.625	2.620	2.638	2.669	2.649
	D_z	0.019	0.264	2.273	4.8423	12.346	18.48	23.10	32.57	51.05	53.74	94.67	255.9	269.0
	$Q(a)$	0.05	0.544	3.076	6.7136	17.741	46.44	58.90	85.07	134.0	140.0	250.0	683.0	712.0
	$r_{\max}$	7.784	5.666	3.589	3.6801	3.821	7.207	7.374	7.586	7.637	7.595	7.692	7.795	7.700
	R	0.0027	0.021	0.068	0.138	0.359	1.257	1.475	1.896	2.546	2.620	3.846	7.795	8.000

注：带下划波浪线（﹏）表示指标性质是错误的，下同。

表 4 窗宽半径 R 取不同值（常数）时的 PPC 建模结果对比

Table 4 Comparison of PPC model in different cutoff radius R (constant)

最优化结果		10^{-5}	10^{-4}	10^{-3}	0.0027	0.068	0.138	0.360	1.260	1.470	1.900	2.500	3.800	7.800	10	100
样本投影值	$z(1)$	−2.536	−0.878	−0.944	−0.387	−0.493	−0.511	−0.521	−0.648	−0.771	−0.918	−1.183	−1.768	−2.359	−2.447	−2.532
	$z(2)$	−1.854	−0.878	−0.944	−0.387	−0.493	−0.520	−0.540	−0.692	0.784	−0.918	−1.107	−1.461	−1.781	−1.820	−1.853
	$z(3)$	−1.535	−0.878	−0.944	−0.388	−0.492	−0.511	−0.517	−0.705	−0.784	−0.918	−1.070	−1.317	−1.511	−1.527	−1.535
	$z(4)$	−0.888	−0.878	−0.944	−0.387	−0.493	−0.511	−0.521	−0.648	−0.692	−0.758	−0.824	−0.905	−0.921	−0.908	−0.889
	$z(5)$	0.045	−0.397	−0.413	−0.387	−0.493	−0.520	−0.539	−0.588	−0.572	−0.562	−0.506	−0.336	−0.075	−0.020	0.041
	$z(6)$	1.520	0.276	0.299	−0.387	−0.493	−0.511	−0.518	−0.315	−0.202	−0.016	0.234	0.729	1.318	1.415	1.515
	$z(7)$	5.249	3.631	3.890	2.323	2.955	3.082	3.156	3.595	3.804	4.091	4.457	5.057	5.329	5.307	5.252
最佳投影向量系数	a_1	0.328	−0.300	−0.165	0	−0.008	−0.006	−0.019	−0.125	−0.12	−0.154	−0.124	0.036	0.239	0.280	0.326
	a_2	0.363	0.163	0.081	0	0.319	0.235	−0.110	−0.061	−0.032	0.012	0.077	0.189	0.319	0.340	0.362
	a_3	0.370	0.423	0.165	0	0.072	0.035	0.036	0.104	0.127	0.156	0.190	0.271	0.351	0.361	0.370
	a_4	0.371	0.417	0.242	0	0.421	0.356	0.301	0.285	0.298	0.313	0.339	0.374	0.380	0.377	0.372
	a_5	0.324	0.423	0.539	0.999	0.747	0.763	0.744	0.700	0.691	0.665	0.634	0.562	0.405	0.369	0.327
	a_6	0.366	−0.231	−0.113	0	−0.074	−0.070	−0.092	0.002	0.026	0.101	0.163	0.250	0.341	0.354	0.366
	a_7	0.338	0.422	0.503	0.033	0.389	0.479	0.572	0.631	0.629	0.616	0.596	0.541	0.410	0.378	0.340
	a_8	0.364	0.347	0.570	−0.009	0.016	−0.014	−0.079	0.037	0.074	0.147	0.210	0.283	0.354	0.360	0.364
目标函数	S_z	2.6732	1.6583	1.7773	1.0246	1.3032	1.3591	1.392	1.591	1.690	1.833	2.026	2.382	2.646	2.665	2.673
	D_z	0.0001	0.0019	0.0190	0.0898	2.5144	4.9535	12.969	42.107	47.134	58.819	74.229	108.06	259.42	365.05	4773.3
	$Q(a)$	1.9×10^{-4}	0.0032	0.0338	0.0921	3.2768	6.7325	18.047	66.974	79.659	107.80	150.38	257.34	686.37	972.93	12760
	r_{max}	7.7844	4.5088	4.8335	2.7110	3.4480	3.6023	3.695	4.299	4.588	5.009	5.640	6.825	7.687	7.754	7.784
	R	10^{-5}	10^{-4}	0.0010	0.0027	0.068	0.138	0.360	1.260	1.470	1.900	2.500	3.800	7.800	10	100

注：带下划线（____）表示投影值很小且几乎为 0。

进一步研究发现，王顺久等（2002）和付强等（2006）推导得出 $r_{max} \leqslant R \leqslant 2p$ 取值范围的过程是错误的。王顺久等（2002）的推导过程可简述为“当 $R \geqslant r_{max}$ 时，有 $0 \leqslant x_{i,j} \leqslant 1$，$|a| \leqslant 1$，因此，$-p \leqslant z(i)=\sum_{j=1}^{p} a_j x_{i,j} \leqslant p$，且 $r_{max} \leqslant 2p$。于是可以得出 R 的合理取值范围为 $r_{max} \leqslant R \leqslant 2p$”。付强等（2006）的推导过程基本类似。显然，以上推导过程至少存在如下两个致命错误：首先，从“$R \geqslant r_{max}$ 且 $r_{max} \leqslant 2p$”，是不可能推导出“$r_{max} \leqslant R \leqslant 2p$”的结果的，用并列式只能推导出 $\begin{cases} R \geqslant r_{max} \\ 2p \geqslant r_{max} \end{cases}$ 的结果，R 值根本不存在 [r_{max}, $2p$] 这样一个闭区间。第二，从“$0 \leqslant x_{i,j} \leqslant 1$，$|a| \leqslant 1$”是不可能推导出“$-p \leqslant z(i)=\sum_{j=1}^{p} a_j x_{i,j} \leqslant p$”的结果的。因为，由 $z(i)=\sum_{j=1}^{p} a_j x_{i,j} \leqslant \sum_{j=1}^{p} a_j$ 和高等数学中的费马（Fermat）极值引理可知，$z(i)$ 取得极大值的必要条件是 $\frac{\partial z}{\partial a_1}=\frac{\partial z}{\partial a_2}\cdots\frac{\partial z}{\partial a_{p-1}}=0$，即只有当 $a_1=a_2=\cdots=a_{p-1}=\sqrt{1-(a_1^2+a_2^2+\cdots+a_{p-1}^2)}$ 时，也即 $a_1=a_2=\cdots=a_p=\frac{1}{\sqrt{p}}$ 时 $z(i)$ 才可能取得极大值，并且 $-\sqrt{p} \leqslant z(i)=\sum_{j=1}^{p} a_j x_{i,j} \leqslant \sum_{j=1}^{p} a_j \leqslant \sqrt{p}$。因此，理论上讲，王顺久等（2002）提出的较大值方案 $r_{max} \leqslant R \leqslant 2p$ 实际上根本是不存在的，必定是错误的。事实上，王顺久等（2002）$R \geqslant r_{max}$ 的假设与 Friedman and Tukey（1974）提出选取 R 合理值时“应使窗宽内的样本点个数既不能太少，以免样本滑动平均时偏差太大，同时又不能随着样本个数 n 的增大而增加太多”的要求也是相矛盾的。因为，当 $R \geqslant r_{max}$ 时，所有样本点毫无疑问都在同一个窗宽内了。而且，用 $-\sqrt{p} \leqslant z(i) \leqslant \sqrt{p}$ 的结论可以很方便地初步判断 PPC 建模结果是否出现了明显的错误，凡是不符合这个结论的，PPC 建模结果必定是错误的，这样的论文有不少（如：张欣莉等，2000；董玉才等，2011）。

3.5　R 在中间适度值（$r_{max}/5 \leqslant R \leqslant r_{max}/3$）范围内取不同值对水质综合评价结果的影响

笔者也建立了 $R=r_{max}/3$ 时的 PPC 模型，其结果如表 3 所示。与 $R=r_{max}/5$ 时的建模结果相比，各个指标的重要性更均衡化，即小的权重都得到了一定程度的提高（如 a_1 从 0.143 到 0.234，a_2 从 0.252 到 0.293），大的权重都有不同程度的下降（如 a_5 从 0.488 到 0.438，a_8 从 0.338 到 0.326），即 Hg、挥发酚和石油类指标对水质的影响程度有所降低，而其他指标的影响程度则都有所提高，致使 DO 较低以及 COD_{Mn} 和 Pb 较高断面的水质评价结果将变差，而 Hg 和挥发酚较高断面的水质评价结果将变好。就 22 个国控断面 110 个样本而言，有 4 个样本的水质类别出现了改变，其中分别有 1 个样本从Ⅳ类变为Ⅲ类和从Ⅲ类变为Ⅱ类，2 个样本从Ⅱ类变为Ⅰ类。

因此，无论 $R=r_{max}/5$ 还是 $R=r_{max}/3$，绝大多数样本（占 96%）的水质类别都保持不变。同时，由于显著提高了 DO 的影响程度，改变了少部分水体的水质类别，但从 PPC 模型输出值是实数的结果来看，改变类别的水质都处于相邻类别水质的分界点附近，对水质实际评价结果的影响并不大。

4　结束语

（1）根据典型国控断面的水质监测指标数据和我国地表水环境质量标准，建立了长江水系水质综合评价的投影寻踪分类（PPC）模型，得到了 2006—2010 年 103 个国控断面的水质综合评价结果。长江水系在国控断面上整体水质较好，Ⅰ～Ⅲ类水质占 92.82%，其中Ⅰ类水质占 62.52%，Ⅱ类水质占 20.58%。流经四川、安徽、江苏等省后，水质明显下降；整体上，上游水质好于下游水质。在化工工业较发达的地区，Hg、石油类、挥发酚和氨氮的污染较严重，出现了劣Ⅴ类水质，水污染现象较严重。在 2007—2009 年期间，水质整体上有明显好转，但 2010 年的水质比 2009 年的水质差，如果任其发展，前期的治理成效将可能前功尽弃。因此，当地政府和有关环保部门，水污染治理工作任重道远，必须采取更严厉的措施，彻底扭转这种恶化趋势，始终保持水污染治理的高压态势，绝不能有丝毫的懈怠和放松。

（2）实证和理论研究都表明，投影窗宽半径 R 值对 PPC 建模结果具有重要影响或者说是决定其结果的唯一参数，目前选取 R 值的较小值方案和较大值方案都是不合理的和错误的。在 R 取很小值（如小于 $0.001S_z$）和很大值（如大于 $2r_{max}$）时，目标函数都只能体现 Friedman 等提出的“使样本投影点整体上尽可能分散”的要求，未能实现“样本点局部上尽可能密集”的目标。而且，当 R 取较小值时，各指标权重的细微改变就可能引起目标函数值的变化，目标函数存在很多局部极小点，致使最优化过程通常都无法求得真正的全局最优解，即使求得了真正的全局最优解，建模结果也没有规律性。R 取中间适度值时，目标函数完整体现了 Friedman 等提出的 PPC 建模的基本思想，而且，在此范围内取不同的 R 值，对建模结果的影响并不大，具有很好的规律性，最优化过程也都能求得真正的全局最优解。R 取常数时也不能得到合理的结果。

（3）采用多种群智能算法对 PPC 模型进行最优化求解，结果表明，GSO、GSA、ABC 和 MGA 的收敛性能较好，收敛速度也较快，应优先推荐使用。

参考文献

陈润羊，花　明，涂安国 . 2008. 长江水系水质评价的几种方法 [J]. *东华理工大学学报 (自然科学版)*, 31(2): 146–151. [Chen R Y, Hua M, Tu A G. 2008. Several methods of water environment quality assessment in the Yangtze River [J]. *Journal of East China Institute of Technology*, 31(2): 146–151.]

董玉才，范格华，张　玲，等 . 2011. 基于投影寻踪法的坦克动力舱热工况综合评价 [J]. *数学的实践与认识*, 41(17): 157–161. [Dong Y C, Fan G H, Zhang L, et al. 2011. Integrated evaluation on thermal working condition in tank power cabin based on projection pursuit method [J]. *Mathematics in Practice and Theory*, 41(17): 157–161.]

封志明，郑海霞，刘宝勤 . 2005. 基于遗传投影寻踪模型的农业水资源利用效率综合评价 [J]. *农业工程学报*, 21(3): 66–70. [Feng Z M, Zheng H X, Liu B Q. 2005. Comprehensive evaluation of agricultural water use efficiency based on genetic projection pursuit model [J]. *Transactions of the CSAE*, 21(3): 66–70.]

付　强，赵小勇 . 2006. 投影寻踪模型原理及其应用 [M]. 北京 : 科学出版社 . [Fu Q, Zhao X Y. 2006. The Principles and Applications of Projection Pursuit Model [M]. Beijing: Science Press.]

金菊良，刘　丽，丁　晶，等 . 2003. 地下水水质评价的逻

辑斯谛曲线模型 [J]. *环境污染与防治* , 25(1): 46–48. [Jin J L, Liu L, Ding J, et al. 2003. Logistic curve model of groundwater quality evaluation [J]. *Environmental Pollution and Control*, 25(1): 46–48.]

李惠明 , 尚广萍 . 1991. 水质现状评价数学模型综合研究 [J]. *中国环境科学* , 11(5): 356–360. [Li H M, Shang G P. 1991. A comprehensive study on the mathematical models of water quality evaluation [J]. *China Environmental Science*, 11(5): 356–360.]

李祚泳 , 郭丽婷 , 欧阳洁 . 2001. 水环境质量评价的普适指数公式 [J]. *环境科学研究* , 14(3): 56–58. [Li Z Y, Guo L T, Ou Y J. 2001. An universal formula suited to water quality evaluation [J]. *Research of Environmental Sciences*, 14(3): 56–58.]

刘小楠 , 崔　巍 . 2009. 主成分分析法在汾河水质评价中的应用 [J]. *中国给水排水* , 25(18): 105–108. [Liu X N, Cui W. 2009. Application of principal component analysis method to assessment of water quality in Fen River [J]. *China Water and Wastewater*, 25(18): 105–108.]

楼文高 , 乔　龙 . 2011. 基于神经网络的金融风险预警模型及其实证研究 [J]. *金融论坛* , (11): 52–61. [Lou W G, Qiao L. 2011. Early warning model of financial risks and empirical study based on neural network [J]. *Finance Forum*, (11): 52–61.]

楼文高 , 乔　龙 . 2013-08-30. 投影寻踪分类建模理论的新探索与实证研究 [J/OL]. *数理统计与管理* , 2015(1), http://www.cnki.net/kcms/detail/11.2242.O1.20130830.1736.001.html. [Lou W G, Qiao L. 2013-08-30. New theory exploration of projection pursuit clustering model and its positive research [J/OL]. *Journal of Applied Statistics and Management*, 2015(1), http://www.cnki.net/kcms/detail/11.2242.O1.20130830.1736.001.html.]

楼文高 , 王延政 . 2003. 基于 BP 网络的水质综合评价模型及其应用 [J]. *环境污染治理技术与设备* , 7(4): 23–26. [Lou W G, Wang Y Z. 2003. Water quality comprehensive assessment model using BP networks and its applications [J]. *Techniques and Equipment for Environmental Pollution Control*, 7(4): 23–26.]

王顺久 , 张欣莉 , 丁　晶 , 等 . 2002. 投影寻踪聚类模型及其应用 [J]. *长江科学院院报* , 19(6): 53–55, 61. [Wang S J, Zhang X L, Ding J, et al. 2002. Projection pursuit cluster model and its application [J]. *Journal of Yangtze River Scientific Research Institute*, 19(6): 53–55, 61.]

张欣莉 , 丁　晶 , 李祚泳 , 等 . 2000. 投影寻踪新算法在水质评价模型中的应用 [J]. *中国环境科学* , 20(2): 187–189. [Zhang X L, Ding J, Li Z Y, et al. 2000. Application of new projection pursuit algorithm in assessing water quality [J]. *China Environmental Science*, 20(2): 187–189.]

Friedman J H, Tukey J W. 1974. A projection pursuit algorithm for exploratory data analysis [J]. *IEEE Transactions on Computers*, 23(9): 881–890.

MacCallum R, Widaman K, Zhang S, et al. 1999. Sample size in factor analysis [J]. *Psychological Method*, 4: 84–99

StatSoft Inc. 2011. Electronic Statistics Textbook [EB]. Tulsa (http://www.statsoft.com/textbook) .

（上接 343 页）

Jia K, Wei X Q, Gu X F, et al. 2014. Land cover classification using Landsat 8 Operational Land Imager data in Beijing, China [J]. *Geocarto International*, doi: 10.1080/10106049.2014.894586.

Lobo F L, Costa M P F, Novo E M L M. 2014. Time-series analysis of Landsat-MSS/TM/OLI images over Amazonian waters impacted by gold mining activities [J]. *Remote Sensing of Environment*, http://dx.doi.org/10.1016/j.rse.2014.04.030.

Michishita R, Jiang Z, Gong P, et al. 2012.Bi-scale analysis of multitemporal land cover fractions for wetland vegetation mapping [J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 72: 1–15.

Niu Z G, Zhang H Y, Wang X W, et al.2012. Mapping Wetland Changes in China between 1978 and 2008 [J]. *China Science Bulletin*, 57(22): 2813–2823.

Wright C, Gallant A. 2007. Improved wetland remote sensing in Yellowstone National Park using classification trees to combine TM imagery and ancillary environmental data [J]. *Remote Sensing of Environment*, 107(4): 582–605.